Identity, Similarity and the OCP

A model of co-occurrence in 107 languages

LabPhon 18

Amanda Doucette (they/them)[1]
Morgan Sonderegger (he/him)[1]
Timothy J. O'Donnell (he/him)[1,2]
Heather Goad (she/her)[1]

[1]McGill University

[2]Mila

June 25, 2022

Introduction

Why is **vowel harmony** so common, and
consonant harmony so rare?

Introduction

Why is **vowel harmony** so common, and
consonant harmony so rare?

To address this question, first we need to know:
Is this actually true, across a large set of languages?

Harmony and Anti-harmony: What we know

Harmony

A tendency for phonologically **similar** segments to co-occur.

- Dozens of studies have identified vowel harmony in individual languages (see Gordon 2016 or Archangeli and Pulleyblank 2007 for a summary).
- Consonant harmony is also attested – although in far fewer languages than vowel harmony (Hansson 2010).
- Ex: Height Harmony
  - [t i t u]
  - *[t e t u]

Harmony and Anti-harmony: What we know

Anti-harmony (Disharmony, OCP, Dissimilation, etc.)

A tendency for phonologically **dissimilar** segments to co-occur.

- Many languages have a restriction against consonants with the same place of articulation (McCarthy 1986; Pozdniakov and Segerer 2007; Graff and Jaeger 2009; Mayer et al. 2010).
- A few languages have productive vowel anti-harmony processes (Harrison 1999; Krämer 1998).
- Ex: Similar Place Avoidance
  - [k a d a]
  - *[t a d a]

Co-occurrence as a Proxy for Harmony

- Harmony and anti-harmony are highly feature-dependent – different languages exhibit harmony with different phonological features.
- As a first step towards modeling harmony and anti-harmony effects, we use an **aggregate similarity measure** to examine rough trends in the data.
  - Pairs that agree in features have higher similarity scores, and pairs that disagree in features have lower similarity scores
  - These are very coarse measures of harmony and anti-harmony, but are sufficient to detect effects in the data.

Research Question 1

Across a large set of languages, is vowel harmony actually more common than consonant harmony? Is consonant anti-harmony more common than vowel anti-harmony?

Hypothesis 1: Yes – similar vowels are more likely to co-occur than similar consonants, and dissimilar consonants are more likely to co-occur than dissimilar vowels.

Research Question 2

Is there any relationship between co-occurrence effects in vowels and consonants?

- **Hypothesis 2** No – dependencies between vowel and consonant effects have not previously been noted.
- **Hypothesis 3** Yes – consonant and vowel co-occurrence effects are correlated.

Research Question 3

Do identical segments display the same co-occurrence patterns as similar segments?

- There is evidence that completely identical segments can behave differently from merely similar segments in some languages (MacEachern 1999; Pozdniakov and Segerer 2007; McCarthy 1986).
- Therefore, the models presented here will account for **identity** effects separately from **similarity** effects.

Similarity Metrics

- Models are fit with **two** different similarity metrics, to ensure results are not driven by a particular way of calculating similarity.
- The same feature set (Panphon) is used for every language (Mortensen et al. 2016).

Feature Similarity

$$FeatSim(x, y) = \frac{|Feats(x) \cap Feats(y)|}{|DistinctFeats|}$$

Natural Class Similarity (Frisch, Pierrehumbert, and Broe 2004)

$$NCSim(x, y) = \frac{|NC(x) \cap NC(y)|}{|NC(x) \cup NC(y)|}$$

Data

NorthEuraLex (Dellert et al. 2020)

- 107 Northern Eurasian languages, 21 families
- IPA transcriptions of 1,016 basic concepts per language

- For each word in each language, we count all co-occurring **consonant** pairs separated by one vowel or diphthong, and all co-occurring **vowel** pairs separated by one consonant (or consonant cluster):
- [kɑnsənənt] would result in the pairs [kn], [sn], [nn], [ɑə], and [əə].

A Bayesian Model of Co-occurrence

- **Previous approaches**: Observed/Expected Ratios (Frisch, Pierrehumbert, and Broe 2004; Walter 2010), Logistic Regression (Graff and Jaeger 2009).
- **Negative Binomial Regression**: A model of *pair counts* – given a set of predictors, how many times is a pair expected to co-occur?
- ***Bayesian* Negative Binomial Regression**: allows us to examine a posterior distribution of model predictions, and simultaneously model consonant and vowel co-occurrences.
  - This allows us to estimate *correlations* between model parameters!

Models

- One model for each similarity metric was fit using *brms* (Bürkner 2017), a front end for the Stan (Stan Development Team 2019) programming language, using weakly informative priors.
- (C_1, C_2) and (V_1, V_2) counts for all pairs in each language are modeled separately, with C and V models tied together across languages.

Models

$$(\mathrm{C/V})\mathrm{PairCount} \sim \mathrm{NegativeBinomial}(p, r)$$

$$\ln\left(\frac{p}{s_1 freq \times s_2 freq}\right) = \beta_0 + \alpha_{lang} + \alpha_{family}$$

$$+\beta_1 \mathrm{sim} + \gamma_{lang}\mathrm{sim} + \gamma_{family}\mathrm{sim}$$

$$+\beta_2 \mathrm{ident} + \delta_{lang}\mathrm{ident} + \delta_{family}\mathrm{ident}$$

$$+\beta_3 \log(\mathrm{Cinv.size}) + \beta_4 \log(\mathrm{Vinv.size})$$

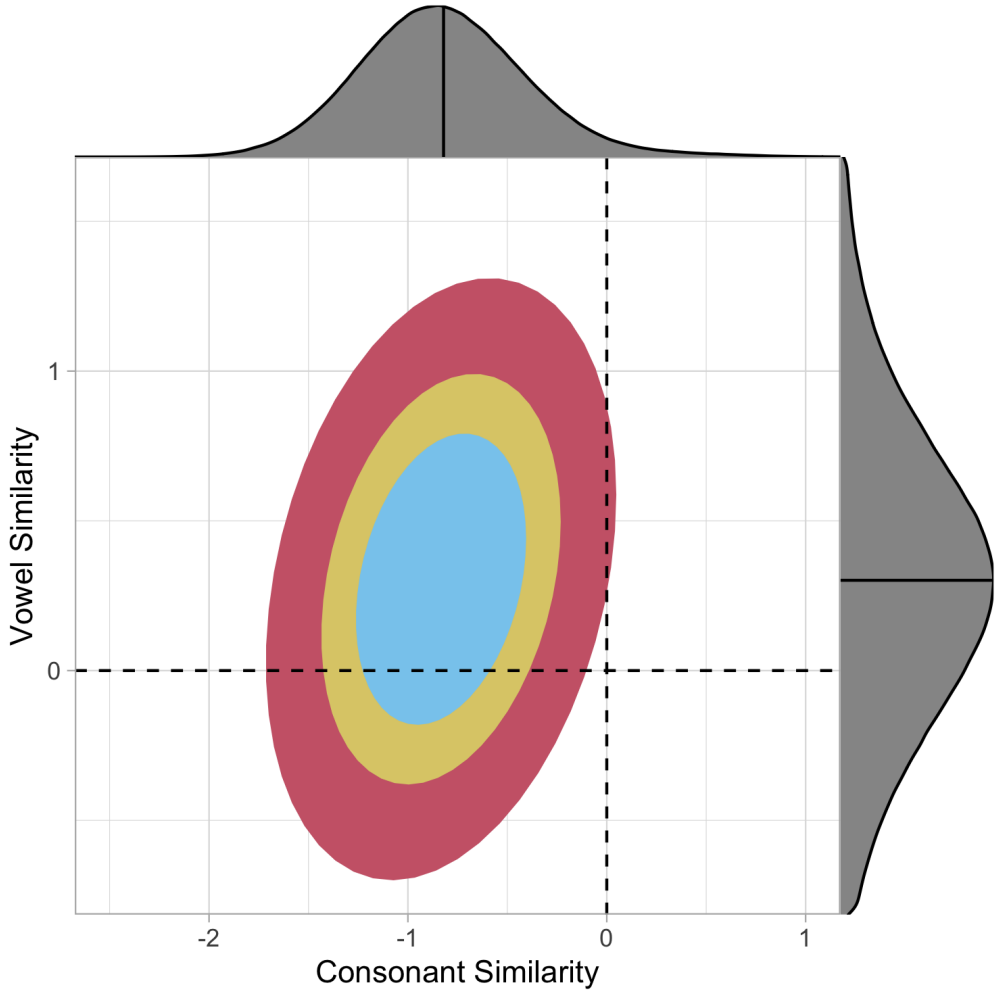
Results: Similarity

- **Dissimilar** Consonants are more likely to co-occur across languages

  β = -0.82, 95% CI [-1.06, -0.59]

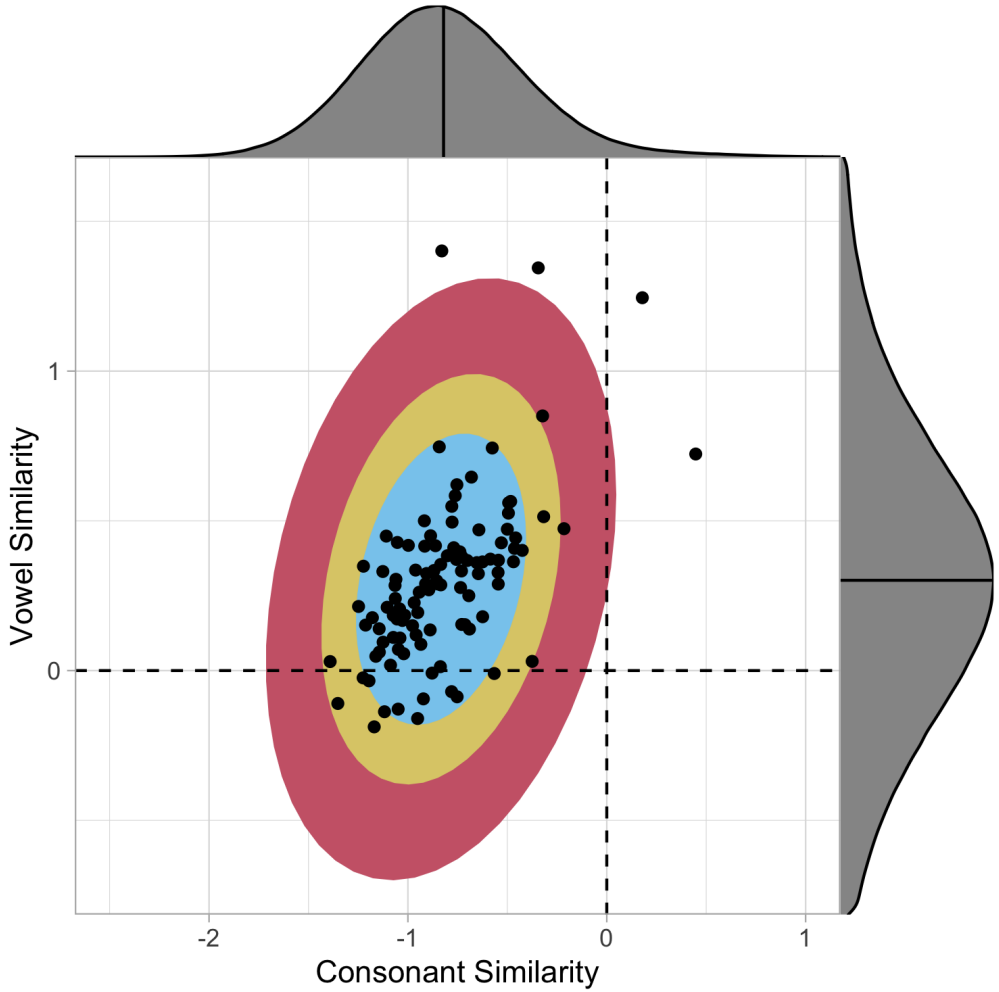
- **Similar** Vowels are more likely to co-occur

  β = 0.30, 95% CI [0.03, 0.58]



Results: Similarity

- Nearly all NorthEuraLex languages have positive Vowel co-occurrence and negative Consonant co-occurrence effects

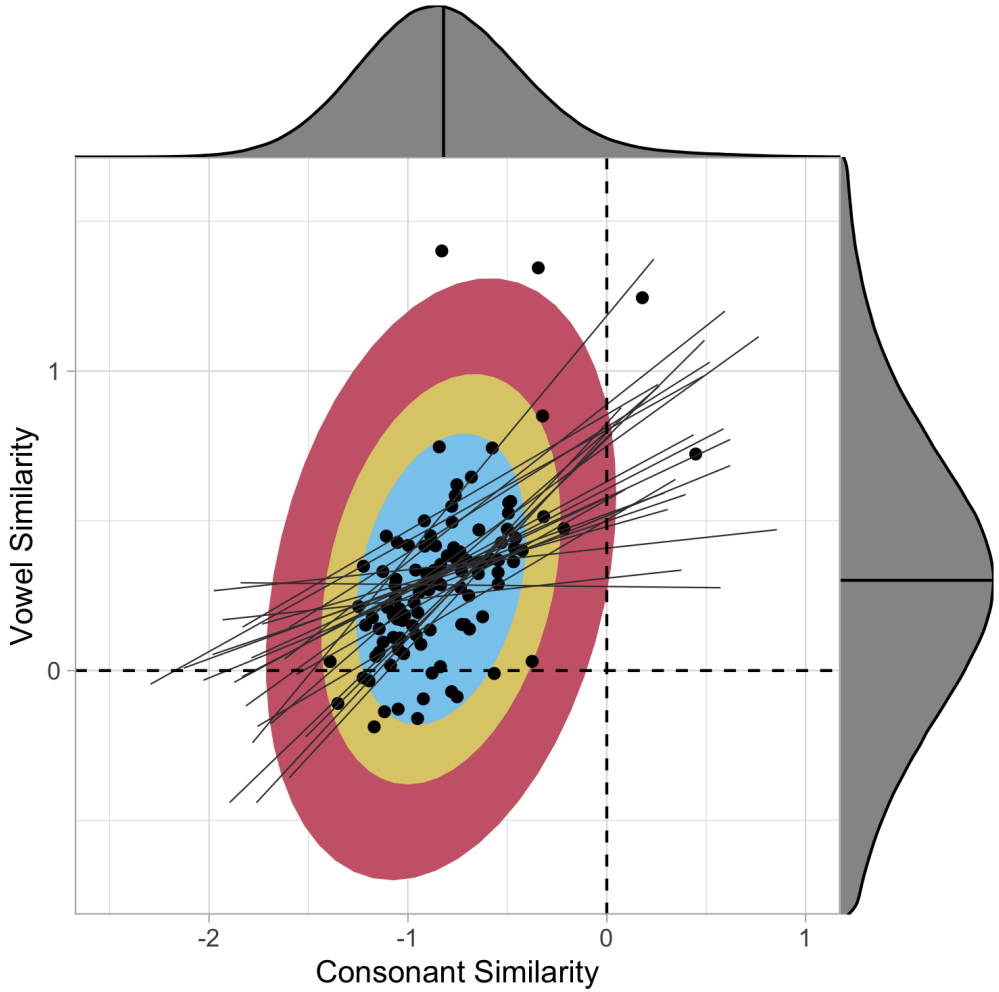


Results: Similarity

- Vowel and Consonant co-occurrence effects are **positively correlated**

  ρ = 0.32, 95% CI [-0.12, 0.70]

- **Stronger** Vowel Harmony effects = **Weaker** Consonant Anti-harmony effects



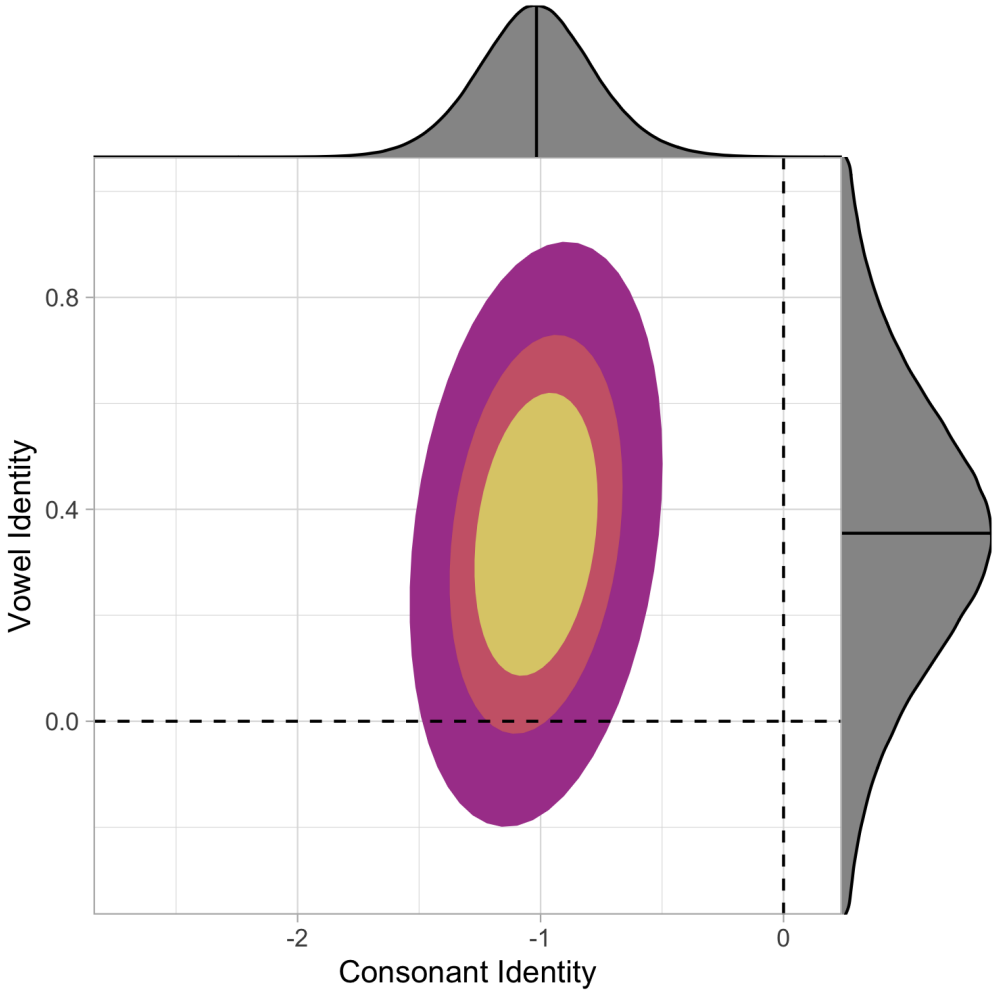
Results: Identity

- Identical Consonants are **less** likely to co-occur

  β = -1.02, 95% CI [-1.23, -0.79]

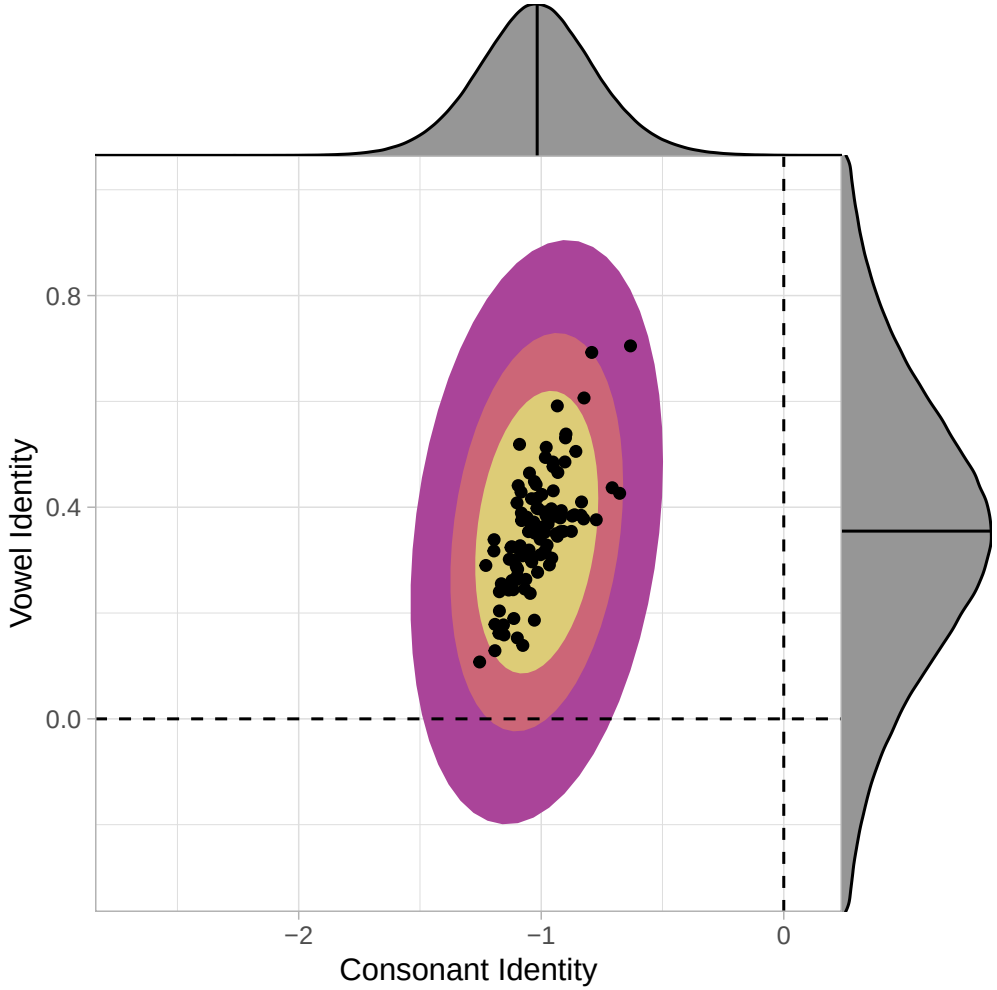
- Identical Vowels are **more** likely to co-occur

  β = 0.35, 95% CI [0.08, 0.62]



Results: Identity

- Nearly all NorthEuraLex languages have negative identical consonant co-occurrence effects, most have positive identical vowel effects

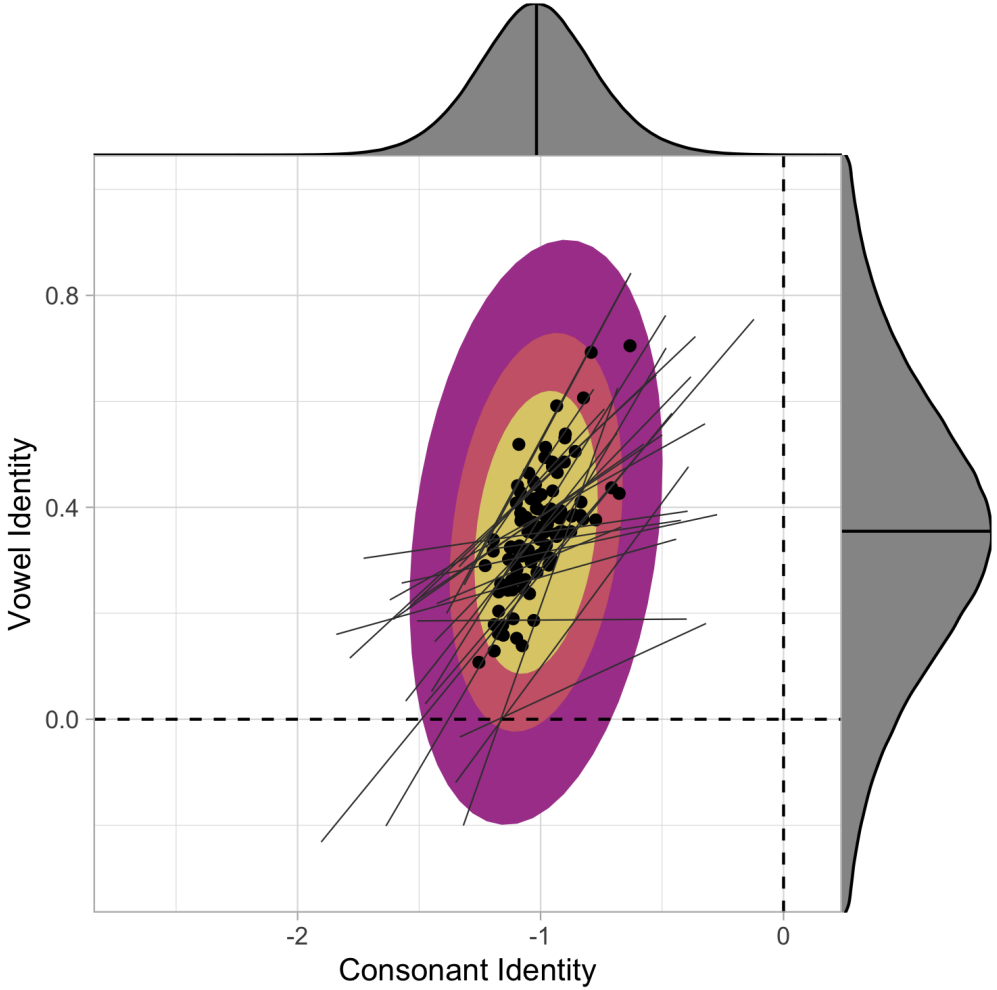


Results: Identity

- Identical Vowel and Consonant co-occurrence effects are **positively correlated**

  ρ = 0.33, 95% CI [-0.26, 0.80]

- **Stronger** Identical Vowel co-occurrence effects = **Weaker** Identical Consonant co-occurrence effects



Research Question 1

Across a large set of languages, is vowel harmony actually more common than consonant harmony? Is consonant anti-harmony more common than vowel anti-harmony?

- **Similar Vowel Co-occurrence** and **Dissimilar Consonant Co-occurrence** are statistical universals across a large sample of languages.
  - This remains true after accounting for individual segment frequency, inventory size, and language family effects.
  - Consistent with vowel harmony and consonant anti-harmony being more common across languages.

Research Question 2

Is there any relationship between co-occurrence effects in vowels and consonants?

- There is a **positive correlation** between consonant and vowel effect strengths – for both Similarity and Identity.
  - This suggests a *trade-off* between consonant anti-harmony and vowel harmony effects.
  - Possible explanation: Harmony reduces a language's ability to encode lexical distinctions, so languages will tend to have strong harmony effects in *either* vowels or consonants.

Research Question 3

Do identical segments display the same co-occurrence patterns as similar segments?

- **Identical** Vowel and Consonant co-occurrence effects follow a similar pattern to similar Vowel and Consonant co-occurrences.
    - All combinations of positive/negative similarity effects and positive/negative identity effects are predicted to be possible, but *[Negative C Similarity, Positive V similarity, Negative C Identity, and Positive V Identity]* is by far the most likely combination.

Discussion

- These effects can be detected with aggregate similarity measures – which features are actually driving the effects?
  - Modeling the effects of individual features, rather than aggregate similarity, would help answer this.
- Although correlations are predicted to be positive, they have large Credible Intervals.
  - Correlations could be between individual features, rather than entire language similarity effects.

Thank you!

References I

- Archangeli, Diana and Douglas Pulleyblank (2007). "Harmony". In: *The Cambridge Handbook of Phonology*. Ed. by Paul de Lacy. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, pp. 353–378. DOI: 10.1017/CBO9780511486371.016.
- Bürkner, Paul-Christian (2017). "brms: An R Package for Bayesian Multilevel Models Using Stan". In: *Journal of Statistical Software* 80.1. version 2.16.3, pp. 1–28. DOI: 10.18637/jss.v080.i01.
- Dellert, Johannes et al. (2020). "NorthEuraLex: a wide-coverage lexical database of Northern Eurasia". In: *Language resources and evaluation* 54.1, pp. 273–301.

References II

- Frisch, Stefan A, Janet B Pierrehumbert, and Michael B Broe (2004). "Similarity avoidance and the OCP". In: *Natural language & linguistic theory* 22.1, pp. 179–228.
- Gordon, Matthew K (2016). *Phonological typology*. Vol. 1. Oxford University Press.
- Graff, Peter and T. Florian Jaeger (2009). "Locality and feature specificity in OCP effects: Evidence from Aymara, Dutch, and Javanese". In: *Proceedings from the annual meeting of the Chicago linguistic society*. Vol. 45. 1. Chicago Linguistic Society, pp. 127–141.
- Hansson, Gunnar Ólafur (2010). *Consonant harmony: Long-distance interactions in phonology*. Vol. 145. Univ of California Press.

References III

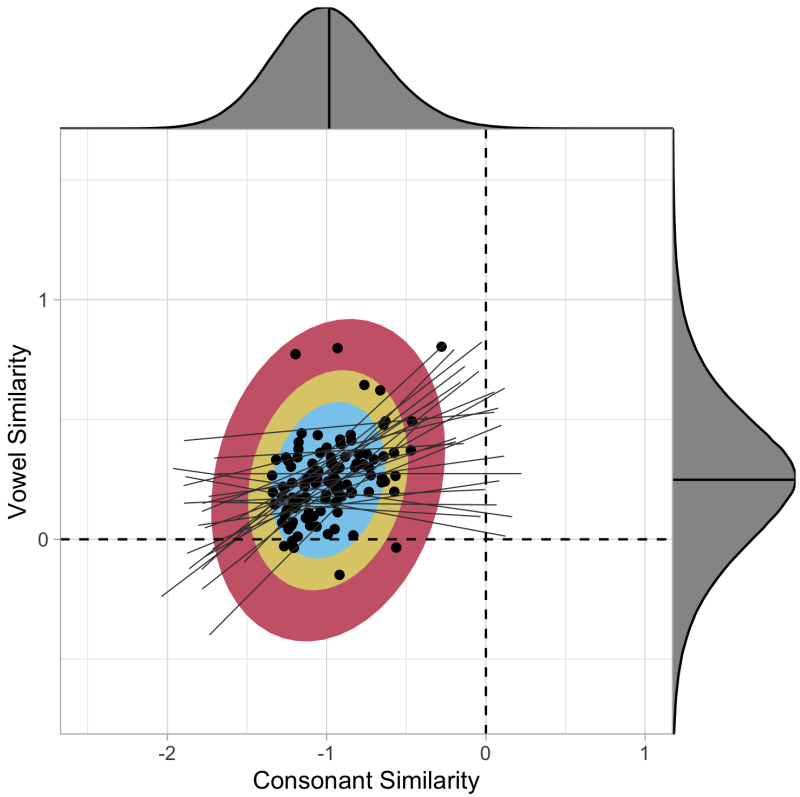
- Harrison, K David (1999). "Vowel harmony and disharmony in Tuvan and Tofa". In: *Proceedings of the Nanzan GLOW*.
- Krämer, Martin (1998). *A correspondence approach to vowel harmony and disharmony*. Sonderforschungsbereich 282.
- MacEachern, Margaret R (1999). *Laryngeal cooccurrence restrictions*. Routledge.
- Mayer, Thomas et al. (2010). "Consonant co-occurrence in stems across languages: Automatic analysis and visualization of a phonotactic constraint". In: *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pp. 70–78.
- McCarthy, John J (1986). "OCP effects: Gemination and antigemination". In: *Linguistic inquiry* 17.2, pp. 207–263.

References IV

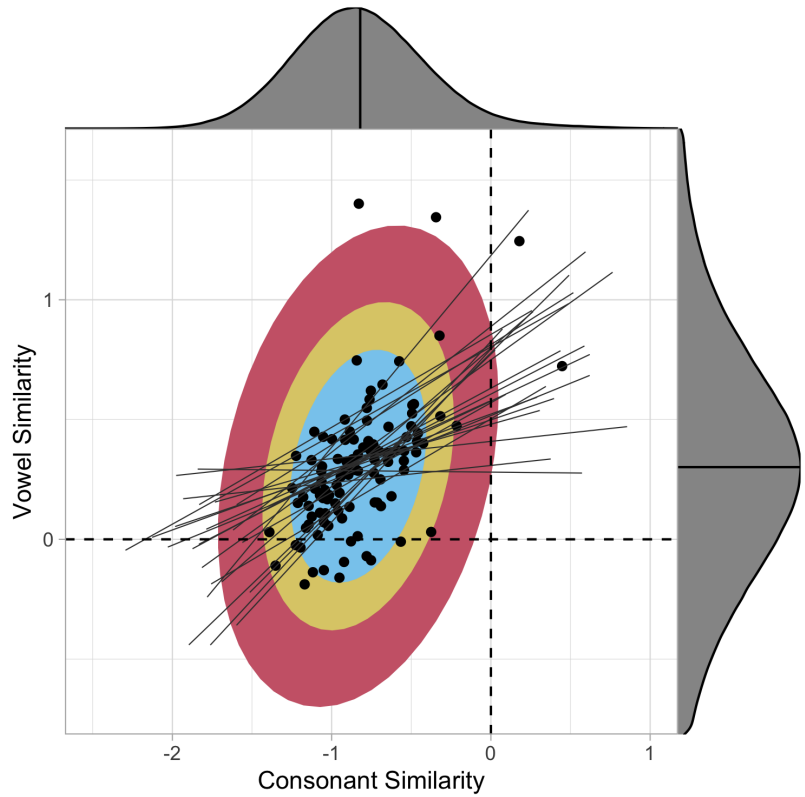
- Mortensen, David R et al. (2016). "Panphon: A resource for mapping IPA segments to articulatory feature vectors". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, pp. 3475–3484.
- Pozdniakov, Konstantin and Guillaume Segerer (2007). "Similar place avoidance: A statistical universal". In: *Linguistic Typology* 11.2, pp. 307–348.
- Stan Development Team (2019). *Stan Modeling Language Users Guide and Reference Manual, Version 2.29*. URL: `http://mc-stan.org/`.
- Walter, Mary Ann (2010). "Harmony versus the OCP: Vowel and Consonant Cooccurrence in the Lexicon". In: *Laboratory Phonology* 1.2, pp. 395–413.

Results: Similarity, Language Family Models

Feature Similarity Model

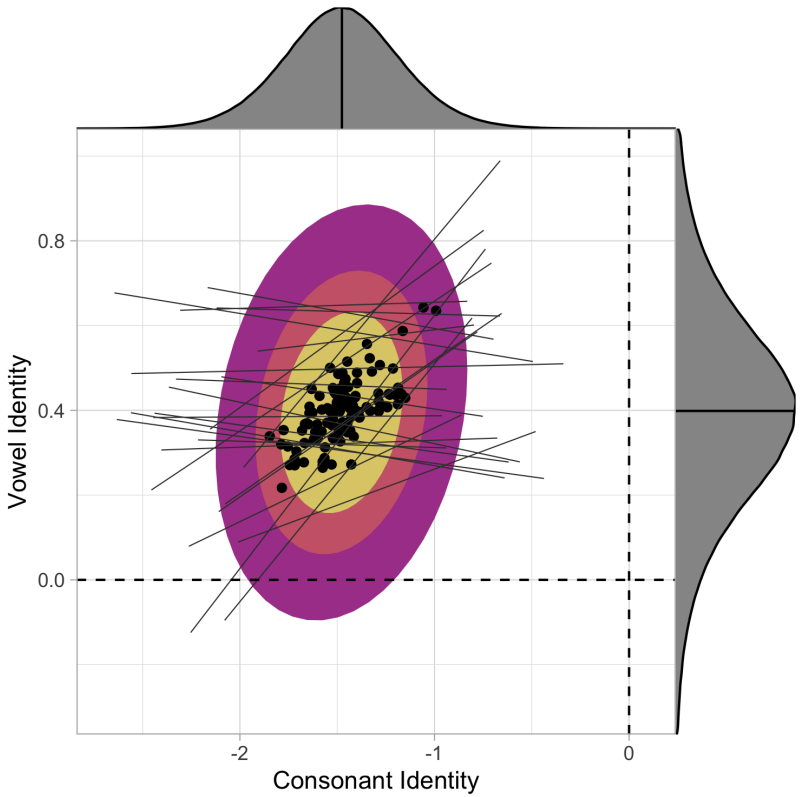


Natural Class Similarity Model

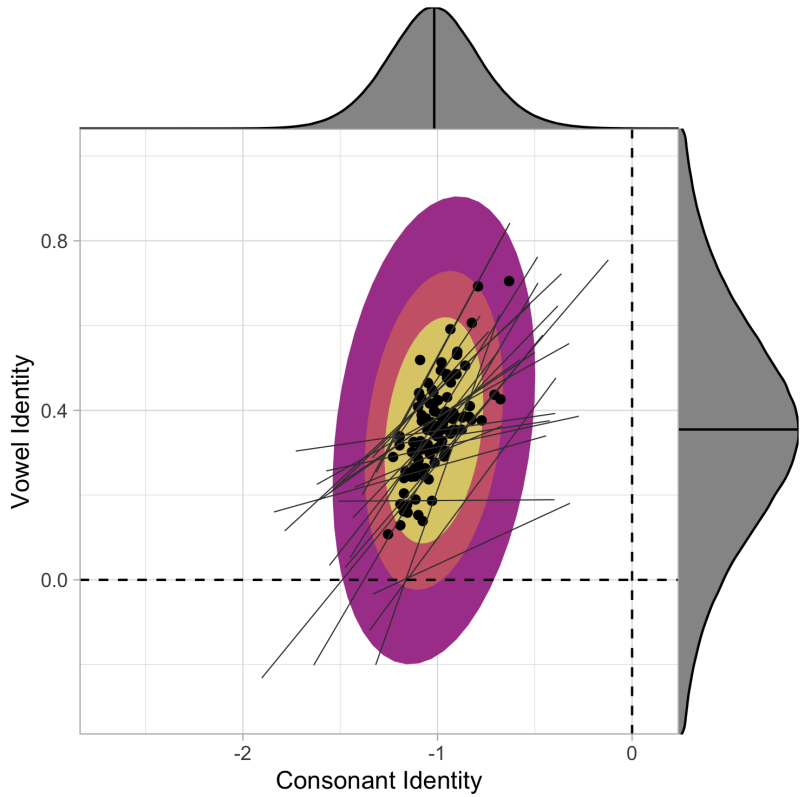


Results: Identity, Language Family Models

Feature Similarity Model

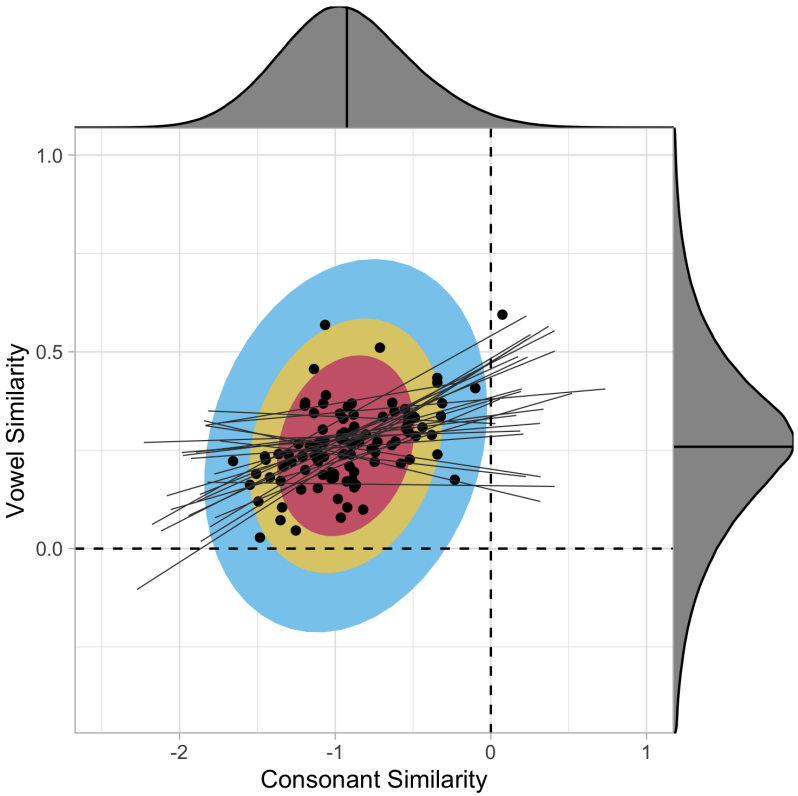


Natural Class Similarity Model

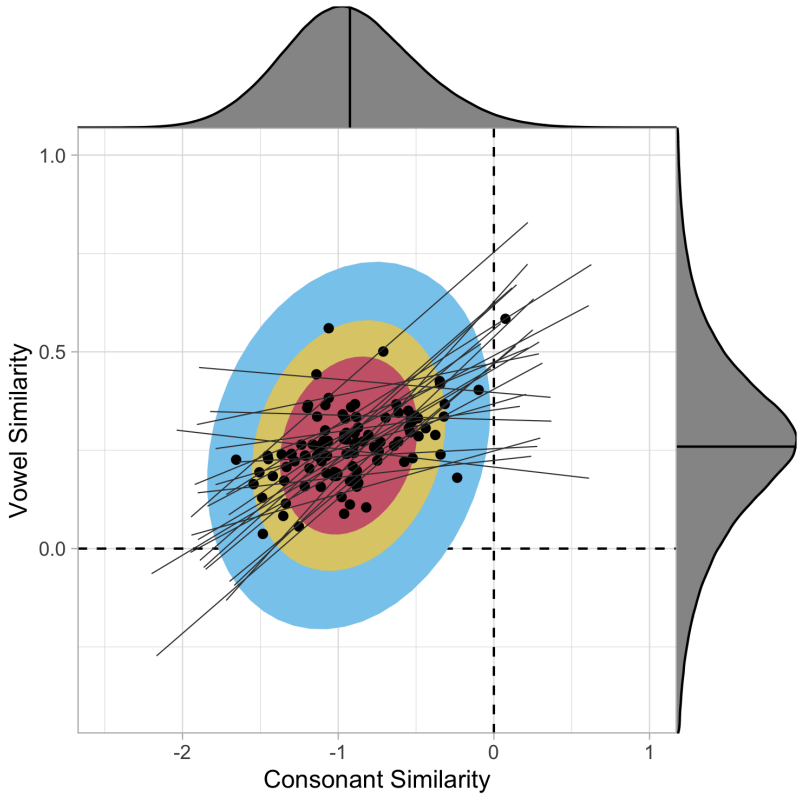


Results: Similarity, no language family

Feature Similarity Model

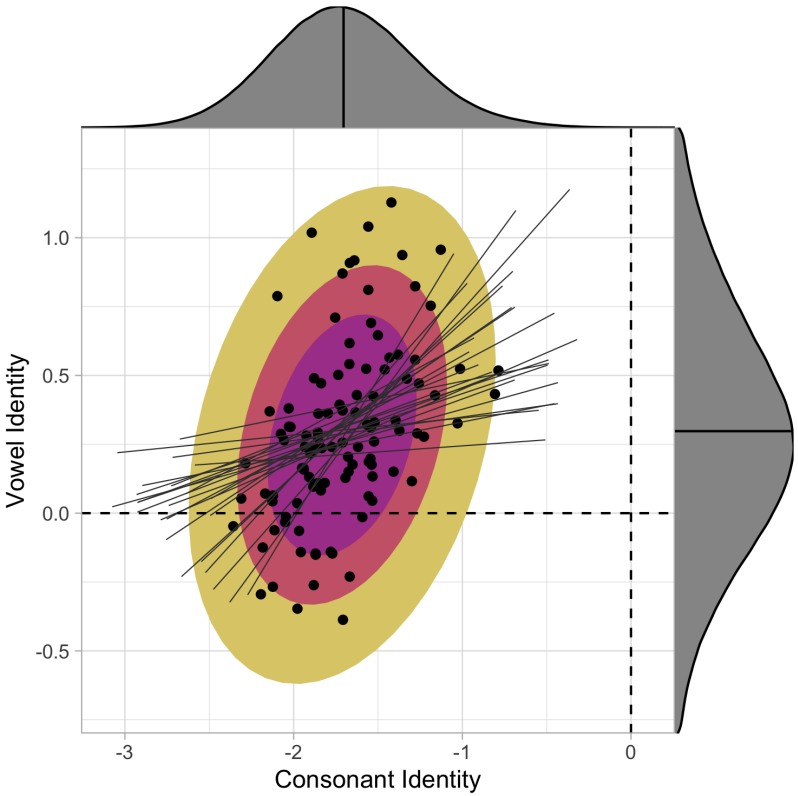


Natural Class Similarity Model



Results: Identity, no language family

Feature Similarity Model



Natural Class Similarity Model